Revista de Investigación Educativa - Vol. 14, n.º 2, págs. 141-163

PRUEBAS ESTANDARIZADAS Y EVALUACIÓN DEL RENDIMIENTO: USOS Y CARACTERÍSTICAS MÉTRICAS

J.M. Jornet Meliá y J.M. Suárez Rodríguez[1]

RESUMEN

En este artículo presentamos algunas reflexiones sobre el uso de las pruebas estandarizadas para la evaluación del rendimiento. Se propone una tipología de pruebas cuyos componentes son: pruebas como Indicadores de resultados, pruebas de certificación y de admisión, pruebas de dominio, pruebas de clase y pruebas individualizadas. Para cada tipo de prueba se revisan las propiedades métricas que se derivan de los objetivos, características y finalidad de las pruebas.

ABSTRACT

In this article we present some reflections about the use of the standarized tests for achievement evaluation. It is proposed a test typology whose components are: tests as outcome indicators, certification and admission tests, domain tests, classroom tests and tailored tests. The measurement properties derived from the tests objectives, characteristics and purpose are reviewed for each kind of test.

1 Dpto. Mètodes d'Investigació i Diagnòstic en Educació. Universitat de València (Estudi General). Avda. Blasco Ibáñez, 21. 46010-Valencia. Tél. y Fax: 96/3864430. E-mail: Jesus.M.Jornet@uv.es / Jesus.M.Rodriguez@uv.es

INTRODUCCIÓN

La utilización de pruebas Estandarizadas en el ámbito Educativo es muy frecuente. En la literatura se utilizan cotidianamente términos que denominan diversos tipos de instrumentos que aluden a este tipo de pruebas: Tests Referidos al Dominio, Tests Referidos a Objetivos, Tests de Competencia, Tests de Certificación, Tests de Dominio, Tests referidos al Criterio, etc.... No obstante, en nuestro medio sociocultural, su uso es más bien escaso y, ciertamente, son pocas las pruebas estandarizadas de Rendimiento que se hayan desarrollado en nuestro país y para nuestro sistema educativo. Las razones que están a la base de este fenómeno pueden ser diversas, pero desde nuestro punto de vista, los usos equívocos de este tipo de pruebas han arraigado la concepción de que son poco útiles a efectos evaluativos y que, en todo caso, su uso está indefectiblemente ligado a corrientes pedagógicas que atienden poco a las características de los individuos. Obviamente, esta posición que atribuimos a buena parte de los detractores de las Pruebas Estandarizadas está simplificada y, probablemente, sería matizada de diversas formas, sin embargo quizá es la posición más generalizada entre ellos.

Desde nuestro punto de vista, el problema normalmente radica en que se pretende de las pruebas estandarizadas usos e interpretaciones para los que normalmente no han sido construidas y, en ocasiones, se desarrollan con esquemas de elaboración que han sido diseñados metodológicamente para objetivos evaluativos diferentes.

Generalmente, la inadecuación de las pruebas está en parte debida a que los criterios de construcción de pruebas se presentan de manera indiferenciada. Aunque son pocos los Modelos de Medida disponibles, las variaciones en su aplicación pueden ser múltiples. En la adaptación precisa de estos elementos radica buena parte de la calidad de las pruebas estandarizadas. Por adaptación nos referimos en este caso al ajuste de los métodos, procedimientos y técnicas de elaboración a las características concretas de la prueba que se desea construir. Estas características son, a su vez, consecuencia del compromiso de diversos factores como son: el objeto de medida, finalidad/uso de la prueba y las características de las personas a las que se desea evaluar a través de la prueba. En este contexto, puede ser de interés realizar algunas reflexiones acerca de los componentes generales de actuación en la elaboración de pruebas estandarizadas, que permitan un mayor aprovechamiento de éstas para los procesos evaluativos.

DIMENSIONES DE CLASIFICACIÓN DE LAS PRUEBAS ESTANDARIZADAS

En la literatura especializada en Medición y Evaluación se identifican una gran cantidad de términos referidos a pruebas estandarizadas. Ante esta diversidad es conveniente determinar algunas dimensiones que nos permitan abordar su clasificación. Entre estas dimensiones nos centraremos en aquéllas que están relacionadas con los componentes métricos o metodológicos de su elaboración. En este caso, el grado de estandarización no entra a formar parte de las dimensiones de clasificación, dado que es una característica constante en todas las pruebas a las que aquí nos referiremos.

Como señalamos en la introducción a este artículo, los tipos de pruebas devienen

de la confluencia de diversos factores. En ellos, podemos identificar dimensiones de definición que afectan a la construcción de las pruebas. Generalmente estas dimensiones son de carácter bipolar y definen un eje en el cuál pueden situarse las características de cada prueba de forma gradual. Así, comentaremos brevemente cada dimensión identificándola con sus polaridades. En todos los casos, al entenderse que estas dimensiones son graduales, cabe establecer un elemento de clasificación intermedio, que corresponden a "enfoques mixtos", que por no ser reiterativos obviaremos en la exposición subsiguiente de dimensiones.

- *Características del Objeto de Medida.* Las características del *Dominio Educativo*[2] a que va dirigido la prueba es un elemento de definición básico que condiciona, desde los elementos de Validez, el conjunto del desarrollo de la prueba. El Dominio Educativo constituye el *Universo de Medida* desde el que se extraen los componentes de la prueba y al que se pretende representar desde ella. Sin entrar en los factores específicos de los tipos de contenidos educativos, las dimensiones a tener en cuenta para orientar el desarrollo de las pruebas son las siguientes:
  1. Amplitud del Dominio Educativo (Dominios amplios vs. reducidos).
  2. Límites del Dominio Educativo (Dominios con limites difusos/no-finitos vs. Dominios con límites concretos/finitos).
  3. Dimensionalidad del Dominio Educativo (Dominios Multidimensionales vs. Unidimensionales).
- *Características de la población a que va dirigida la prueba.* Afecta fundamentalmente a la elección del Modelo de Medida y la selección de indicadores que permitan el análisis adecuado del ajuste de las características de la prueba a las de la población. Las dimensiones más relevantes son:
  4. Amplitud de la población (Población extensa vs. Reducida).
  5. Grado de diversidad de la población (Población Heterogénea vs. Homogénea).
- *Finalidad y uso de la prueba.* La Validez no es en sí misma una característica imputable a una prueba, es más bien el uso que se pretende realizar de las puntuaciones de ella derivadas lo que debe analizarse como elemento de validación (Hambleton; 1984). Así, la utilización que se desee realizar de la prueba tiene consecuencias desde la definición del Dominio Educativo hasta el establecimiento de Estándares de puntuación. Las dimensiones más importantes que pueden identificarse en este punto son:
  6. Decisiones asociadas al uso de la prueba (De carácter Formativo vs. Sumativo).
  7. Unidades sobre las que se pretenden tomar decisiones (Individuos vs. Grupos).
- *Características del tipo de Interpretación de puntuaciones.* Como en el caso anterior, estos elementos afectan a todo el desarrollo de la prueba. La dimensión central a que pueden reducirse estas características es:
  8. Tipo de Estándar de referencia (Normativo vs. Criterial).

---

2 Por Dominio Educativo nos referimos al conjunto de objetivos, contenidos, actividades y tareas que constituyen el objeto de la educación, sea en general sea en un programa concreto (Jornet y Suárez, 1989a).

El cruce de estas dimensiones puede servir para identificar las características de diversos tipos de pruebas estandarizadas y orientar los componentes específicos de sus procesos de elaboración. Una propuesta, aunque no exhaustiva, de tipología de pruebas estandarizadas se recoge en el Cuadro 1. Los elementos metodológicos y los aspectos que entendemos más relevantes en su construcción los comentaremos a continuación.

PRUEBAS DE AMPLIO ESPECTRO

En este apartado revisamos las características y usos de pruebas de Rendimiento que se orientan a la evaluación de grandes Áreas o Dominios Educativos. En esta categoría incluimos las pruebas que se utilizan como Indicadores de Resultados para

CUADRO 1
TIPOS FUNDAMENTALES DE PRUEBAS ESTANDARIZADAS VALORADOS RESPECTO A OCHO DIMENSIONES BÁSICAS DE CARACTERIZACIÓN

	Dimensiones de Valoración							
Tipo de Prueba Estandarizada	Amplitud del Dominio Educativo	Límites del Dominio Educativo	Dimensionalidad[3]	Amplitud de la Población	Grado de diversidad de la Población	Decisiones Asociadas	Unidades sobre las que se decide	Tipo de Estándar
De Amplio Espectro: • Indicadores De Resultados • Certificación • Admisión	Amplio	No-finitos Difusos	Multidimensionales	Amplia / Muy amplia	Heterogénea	Formativas/ Sumativas Sumativas	Grupos Individuos	Mixto: Normativo Normativo Criterial
De Nivel o Dominio	Mixto	Mixto	Multidimensionales	Intermedia / Amplia	Mixto	Sumativas	Individuos	Criterial
De Clase	Reducido	Finitos Concretos	Unidimensionales	Reducida / Muy reducida	Homogénea	Formativas/ Sumativas	Individuos	Criterial
De Propósito Diagnóstico	Mixto	Finitos Concretos	Mixtos	Amplia/ Reducida	Heterogénea	Formativas	Individuos	Normativo Criterial
Individualizadas	Reducido	Finitos Concretos	Unidimensionales	Reducida	Homogénea	Formativas/ Sumativas	Individuos y/o Grupos	Criterial

3 Hace referencia a las características originales del Dominio Educativo. Todas las pruebas es preciso adecuarlas a Universos Unidimensionales, por lo que en el caso de universos multidimensionales, se focalizan las pruebas sobre regiones específicas del Dominio.

el Análisis y/o Evaluaciones de Sistemas Educativos, Centros y Programas, pruebas de Certificación y pruebas de Admisión.

- ***Pruebas Estandarizadas como Indicadores de Resultados.***

La actividad evaluativa forma parte de la cultura de gestión de los Estados democráticos. En el ámbito pedagógico pueden observarse diversos modelos y enfoques de Evaluación de los sistemas educativos en los que confluyen indicadores de diferente índole.

Para la construcción de indicadores de resultados, parece claro que cuando se aborda el análisis de un Sistema Educativo, de un Programa o de un Centro, uno de los indicadores a tener en cuenta —aunque no de forma exclusiva— son los resultados esenciales del programa (De Miguel, et al., 1994; Pérez Juste y Martínez Aragón, 1989; Tejedor et al., 1994).

En estos contextos es necesario utilizar pruebas estandarizadas que "traduzcan" los niveles de competencia que en las diferentes disciplinas y materias, una sociedad asume como objetivo educativo.

En este sentido, no es posible abordar un análisis adecuado de un sistema o un Programa si no se cuenta con pruebas estandarizadas de probada fiabilidad y validez.

Así, buena parte de los modelos de evaluación de Sistemas Educativos basados en indicadores[4] incorporan indicadores de resultados del aprendizaje de los alumnos sustentados sobre pruebas estandarizadas, diferenciándolos de las calificaciones escolares o de otros indicadores de síntesis (como las tasas de egresados) que suelen identificarse como Resultados del Sistema. En los campos de la evaluación de centros y de programas también resulta habitual la utilización de estas pruebas como indicadores. Mayor tradición, si cabe, tiene la utilización de pruebas de este tipo como indicadores para actuaciones evaluativas a la medida en muy diversos niveles educativos, ámbitos de referencia y objetivos (a partir de los servicios de instituciones como el ETS en USA, el APU para Inglaterra, Gales e Irlanda del Norte; o el CITO en el contexto holandés-alemán).

¿Qué componentes están implicados en la elaboración de estas pruebas?

La definición del Dominio a que se refieren estas pruebas debe realizarse por un Comité de Expertos en la Materia objeto de evaluación, apoyados por especialistas en Medición y Evaluación como asesores metodológicos. Los problemas que deben enfrentar este tipo de Comités son variados y de su adecuada solución depende en

4 Existen sistemas de indicadores que permiten la comparabilidad entre diversos países y utilizan un número relativamente reducido de pruebas como es el caso de la OCDE (CERI/INES; 1995) o la Comunidad Europea (West et al., 1995). A un nivel intermedio se encuentran los programas desarrollados por la IEA (Postlethwaite, 1987), finalmente existen otros más completos en cuanto a la información que emplean sobre productos educativos como el sistema federal USA (SSPEI, 1991) o algunos otros sistemas más recientes que se están impulsando en el ámbito iberoamericano (por ejemplo la propuesta de Martínez Rizo; 1996). En nuestro país, el Instituto Nacional de Calidad y Evaluación (INCE) está desarrollando pruebas como indicadores de resultados del sistema educativo, habiéndose comenzado a publicar los primeros resultados (Gil, González y Suárez, 1995; INCE, 1996).

buena medida la validez y utilidad de las pruebas resultantes. Pasamos a revisar brevemente los elementos que caracterizan métricamente el desarrollo de estas pruebas y, en consecuencia, condicionan su uso.

El *Dominio Educativo*, como Universo de Medida, suele ser muy amplio, referido generalmente a una materia o disciplina considerada en función de los objetivos terminales de todo un período educativo (por ejemplo, las Matemáticas o el Lenguaje en la Primaria, o al final de la Secundaria) o, a lo sumo, se focalizan sobre grandes dimensiones de las mismas (Medida o Algebra, Comprensión Lectora...). Además, los limites del Dominio suelen ser difusos, dado que junto a la amplitud del mismo, se suma como dificultad añadida en la definición del Dominio el hecho de que suelen ser pruebas cuya finalidad es la evaluación en una gran población de sujetos, afectados por muy diversas aplicaciones de los Diseños Curriculares de referencia, desarrollados sobre diferentes modelos didácticos.

Este hecho conlleva que nos encontremos ante Dominios en la práctica no-finitos, en los que no es posible establecer una estrategia de muestreo probabilístico desde el Universo de Medida para configurar la Prueba[5]. De esta forma, la definición del Dominio debe realizarse sobre productos educativos esenciales, equiprobables a través de diferentes subpoblaciones y que mantengan sus parámetros fundamentales invariantes en las subpoblaciones identificables en la población.

Una dificultad adicional en la definición de este tipo de Dominios reside en que, por la amplitud del Universo de Medida, generalmente no son unidimensionales y están apoyados en constructos teóricos débiles[6], con pocas evidencias de validación. En este tipo de pruebas, los avances mas claros corresponden a estructuraciones dimensionales por el nivel cognitivo que implican las tareas-ítems.

Así, si bien la base de desarrollo de la definición del Dominio recae sobre el juicio de expertos, también es cierto que su comprobación se sustenta fundamentalmente sobre comprobaciones estructurales basadas en el análisis empírico de los resultados de las pruebas. En este sentido, un problema adicional que encontramos en estos desarrollos es que la comprobación de la Unidimensionalidad[7] es difícil, pues, aunque existan propuestas metodológicas específicas para este tipo de análisis con variables dicotómicas, ciertamente los resultados son insatisfactorios dado que las dimensiones en muchas ocasiones se identifican por la dificultad de los ítems y no son interpretables desde los contenidos.

Por ello, en las estrategias de desarrollo de este tipo de pruebas es especialmente

---

5 Ante la imposibilidad de concretar todos los miembros del Universo de Medida, se pasa a utilizar estrategias de representación fundamentadas en tipologías básicas o elementos clave en la estructura del Dominio.

6 Es decir, no suelen estar desarrollados a partir de una teoría de aprendizaje que, de manera holísta, globalice y de sentido a la estructuración y funcionalidad de la prueba.

7 Siendo este un supuesto básico sobre el que se sustenta la medida desde cualquiera de los modelos actualmente existentes y que se enraiza en la información que se extrae en cada elemento de la prueba (Hambleton y Swaminathan, 1985; Osterlind, 1992). Aunque se han planteado algunas alternativas para superar este problema, como los trabajos de Reckase (1979) o Samejima (1974), hasta el momento no se pueden considerar como auténticas opciones disponibles.

importante el trabajo realizado por el Comité de expertos, anticipando la estructura teórica posible del Dominio e identificando regiones en el mismo, de forma que para cada uno de los subdominios se configuren pruebas específicas.

Junto a los problemas teóricos contemplados en líneas anteriores, es preciso considerar en la definición del Dominio de estas pruebas, algunos aspectos de orden práctico ligados a la funcionalidad de las mismas. Si se desea desarrollar pruebas que sirvan como Indicador de Resultados y utilizarlas en evaluaciones sucesivas, no es útil configurar una sola prueba[8]. La alternativa de elaborar Formas Paralelas es compleja y difícil de lograr. En este marco, la creación de Bancos de Reglas de Generación de Elementos[9] —lo más deseable— o Bancos de Ítems —lo más frecuente— se configuran como alternativas que nos acercan a la posibilidad de disponer de pruebas aleatoriamente paralelas o al menos equivalentes.

En este contexto de desarrollo es especialmente importante el trabajo que realiza el Comité de expertos en cuanto a la formulación y revisión de Ítems. Así, un aspecto crucial en la elaboración de este tipo de pruebas es el Análisis Lógico de Ítems que se centra sobre diferentes elementos.

En primer lugar, respecto a la selección inicial de ítems, es conveniente basarla en dos dimensiones del contenido: a) la importancia de los ítems, y b) su dificultad teórica. Esta consideración de dos dimensiones facilita que los ítems sean propuestos y seleccionados desde la idea, antes señalada, de que que representen conocimientos o habilidades esenciales, cubriendo a su vez diferentes estratos de dificultad. Desde esta estrategia se evita que la selección de ítems se contamine con la idea de "mínima competencia", la cual, en ocasiones, es entendida como expresión de la dificultad —y no de la importancia—: este hecho constituye una desviación frecuente del trabajo de estos comités.

Otro elemento a tener en cuenta en la formulación de los ítems es su calidad técnica, la cual es necesario revisar inicialmente por procedimientos lógicos y, posteriormente, basándose en resultados de ensayos piloto. En el caso de utilización de ítems de Elección Múltiple, el análisis de distractores debe constituir un trabajo central de este aspecto. Así, es tanto más importante el control y la anticipación de la dificultad y la adivinación desde la formulación de los ítems, que desde el análisis empírico de resultados.

Junto a estos elementos, el análisis del Sesgo supone un aspecto clave para la validez de este tipo de pruebas. Debe tenerse en cuenta que éstas son pruebas destinadas a la Medición y Evaluación de un Dominio en una amplia población de referencia. De esta manera, es conveniente anticipar en el análisis lógico las variables que podrían

---

8 Una sola aplicación de una prueba de estas características puede inhabilitar su uso. Una vez es conocida una prueba de este tipo puede ser utilizada como objeto directo de aprendizaje.

9 Son procedimientos que concretan de forma unívoca al ítem de forma que su escritura se torna automática, entre ellos los más destacados son los que se recogen en Roid y Haladyna (1982). Aunque lentamente, los procedimientos han ido evolucionando para tratar de hacer frente a la evaluación de la actuación en tareas complejas, incrementando el nivel cognitivo de la evaluación (por ejemplo, los conjuntos de ítems —ítem sets— desarrollados por Haladyna —1992—, o el modelado de ítems desarrollado por La Duca —La Duca et al. 1986—).

actuar como fuente de sesgo. No vamos a extendernos aquí en estos aspectos, dado que han sido anteriormente expuestos en esta Revista (Jornet y Suárez, 1990; Ordeñana, 1991); sin embargo, en este tipo de pruebas es básica la independización del sesgo, el cual puede provenir de diferentes fuentes como el Sexo, el ámbito territorial, los niveles socioculturales o socioeconómicos, la Lengua, etc...

El control de todos estos elementos se basa en dinámicas de juicio bien establecidas, en las que en la síntesis de resultados se utilizan procedimientos de análisis de la consistencia inter-jueces.

En este sentido, hay que destacar la necesidad de los procedimientos de juicio, el estudio de las dinámicas más adecuadas a seguir por parte de los Comités y de los tipos de análisis a realizar, tanto como elementos de síntesis de la opinión de los Comités, como para detectar jueces que aportan valoraciones extremas, etc.

La Validez de las pruebas se asienta sobre procesos de análisis teórico de los componentes de medida y la revisión lógica de todas las unidades[10]. Es fundamental reconocer el valor de los procesos de juicio en este ámbito, los cuales deberán ser atendidos adecuadamente. No obstante, un problema habitual que se contempla en este tipo de pruebas es que muchas de ellas se sustentan más sobre el Modelo de Medida utilizado que sobre el análisis teórico del Dominio Educativo. Por mencionar tan sólo uno de los elementos clave en que se asienta la Validez de Constructo. De hecho, el problema estriba en que los principios que se refieren a la validez y que están recogidos en cualquiera de los modelos de medida no son sino una parte relativamente reducida de los indicios que definimos como facetas de la Validez de Constructo (Angoff, 1988). Por ello las aportaciones de los modelos de medida deben entenderse en un plano instrumental dentro de la estrategia global de validación y no a la inversa.

¿Qué Modelo de Medida es más adecuado en este contexto?

Si se pretende utilizar estas pruebas sobre una población amplia, la base métrica necesariamente se encuentra en la Teoría de Respuesta al ítem que favorece el desarrollo de pruebas sobre parámetros invariantes de los ítems y que permiten una graduación adecuada de los ítems asociados con la habilidad general que mide la prueba (Hambleton y Swaminathan, 1985; Weiss y Yoes, 1991). No obstante, estos modelos han demostrado su adecuación, hasta el momento, con dos condiciones bastante precisas no existiendo un acuerdo generalizado sobre su utilización cuando alguna de ellas no se cumple. La primera condición resulta de la unidimensionalidad del constructo y la segunda del tamaño de la población referente para establecer la invarianza (Linn, 1990; Osterlind, 1992).

Un elemento adicional, que guía la selección de Indicadores tanto para el Análisis de Ítems, como para la Fiabilidad, es el tipo de interpretación de las puntuaciones que se requiere. Así, la dicotomía Normativa-Criterial está a la base de esta selección.

Las pruebas estandarizadas de Rendimiento como Indicadores de Resultados no son interpretadas individualmente, por lo que, en principio, no sería necesario ningún elemento que coadyuve a la interpretación. Así, respecto a la selección de indicadores,

10 Y debe ser refrendada por la acumulación de evidencias empíricas, tanto desde una base de investigación experimental como correlacional.

el Análisis de parámetros de los ítems y el sesgo, junto a comprobaciones de la fiabilidad como consistencia global serían suficientes para un planteamiento métrico adecuado de las pruebas.

Sin embargo, el contexto de uso de estas pruebas —aunque no requieran de un Estándar para la interpretación individual de puntuaciones— lleva a que se necesite algún procedimiento global. De este modo, estas pruebas se utilizan en Evaluaciones sobre la Calidad de un Sistema, un Centro o un Programa y, por ello, debe tenerse en cuenta que de no acompañarse de ningún elemento interpretativo la Evaluación quedará en el terreno meramente descriptivo.

En éste ámbito, las informaciones normativas son indudablemente la base de análisis y la referencia más clara respecto a las características del Dominio evaluado. No obstante, es conveniente que el Comité de Expertos que desarrolla el Análisis y Especificación del Dominio establezca además un Estándar[11] —basado en juicio— que identifique, al menos, los niveles mínimos de competencia aceptables como indicador de suficiencia del sistema. El Estándar operativiza la idea de calidad. Este tipo de Estándares, son especialmente útiles en estudios Longitudinales, pues es conocido que cuando trabajamos con grandes muestras, pequeñas diferencias se identifican como diferencias estadísticamente significativas, y es necesario algún tipo de criterio que nos permita valorar la "cualidad de la diferencia".

Para el desarrollo de este tipo de Estándares se puede trabajar desde Metodologías específicas de Estándares basados en los ítems. Son especialmente de interés para este tipo de pruebas los procedimientos desarrollados para situaciones multivariadas de decisión, como por ejemplo las propuestas de modificación del método de Angoff (Hambleton y Plake, 1995), el procedimiento de Jaeger (1993) o la síntesis formulada por Putnam, Pence y Jaeger (1995). Estos procedimientos abordan la toma de decisiones en tareas complejas, partiendo de la base de que lo que se busca es un perfil de ejecución a través de un conjunto de dimensiones relevantes; lo que es el caso de una buena parte de las situaciones que se encuentran dentro de este tipo de pruebas. En definitiva, constituyen un avance en la linea de operativizar la idea de calidad a partir de los contenidos evaluados, reteniendo la complejidad consustancial a la magnitud del Dominio a que se refieren este tipo de pruebas.

• ***Pruebas Estandarizadas de Certificación y de Admisión.***

Estas pruebas tienen por objeto recoger la información que permita certificar que una persona ha superado administrativamente un determinado nivel educativo o que tiene los conocimientos necesarios para ser admitido en un programa de formación de amplio espectro, por ejemplo, en la enseñanza universitaria[12]. Son pruebas que, por

11 Entendemos por Estándar la puntuación en el Dominio que indica el nivel mínimo de competencia. Se trata de la expresión de este nivel en la Escala de puntajes verdaderos, teóricos, libres de error (Jornet y Suárez, 1989b).

12 En nuestro país existen pocas experiencias aún desarrolladas con este tipo de pruebas. Estudios de interés a este respecto son los de Tourón (1985) y Toca y Tourón (1989), en el ámbito universitario, o en relación a la Educación General Básica los trabajos llevados a cabo por Rivas et al. (1986), que desarrollaron una línea de investigación que condujo a la elaboración de pruebas para los finales de Ciclo en la EGB en diversas materias.

tanto, se orientan a un universo instruccional muy amplio, cuya definición es básicamente empírica —es decir, muy operativizada—, y que se centran muy especialmente en el producto educativo.

Así, si se pretende establecer un nivel generalizado que certifique unos conocimientos mínimos para superar la Secundaria Obligatoria, obviamente estamos hablando de abordar la medición y evaluación de un Dominio educativo que se extiende a lo largo de cuatro años y que está concretado en un número importante de materias diferenciadas. Además, parece razonable tener presente que para esta situación no existen aportaciones teóricas que nos permitan extraer conclusiones ciñéndonos a la valoración de unas cuantas dimensiones. Como mucho, podremos efectuar una definición bastante pormenorizada del Universo de Medida eligiendo algún punto de referencia como pueden ser los textos legales que reflejan las orientaciones y objetivos necesarios en estos niveles. En relación directa con esta cuestión, si no se dispone de un marco teórico sólido de referencia y se debe abarcar un universo muy amplio es básicamente imposible abordar una evaluación del proceso, por lo que estas pruebas se suelen concentrar en la valoración del producto educativo. Las reglas de conexión entre el Universo de Medida de referencia y la prueba concreta no se pueden especificar de forma exhaustiva, por los mismos motivos que acabamos de apuntar.

Las referencias a la definición del Dominio Educativo, que señalamos para las pruebas anteriores, son aplicables aquí. Únicamente debe tenerse en cuenta que en la selección de unidades del Dominio prevalecerán los juicios acerca de la relevancia de los ítems como expresión de competencia, dado que ello es especialmente importante para poder establecer el Estándar.

• ***¿Qué tipo de interpretación se requiere en estas pruebas?***

A diferencia de las pruebas descritas anteriormente, hay que considerar que a partir de estas pruebas se pretende realizar una interpretación específica del nivel de competencia de cada persona, por lo que el planteamiento global de desarrollo de las pruebas varía sustancialmente en la selección de indicadores. Precisamente debido a esta referencia individual en el objetivo de valoración, en este contexto, cuando se trata de la valoración de personas pertenecientes a una población muy amplia resulta especialmente importante enfatizar el análisis del sesgo para asegurar la equidad de la evaluación.

El establecimiento del estándar de superación es difícil que se refiera exclusivamente a un valor absoluto. En este sentido, hay que tener presente que la propia amplitud y heterogeneidad del contenido hacen muy difícil poder definir exactamente cuál es el nivel mínimo exigible —mediante objetivos o conocimientos específicos— a un sujeto para alcanzar el nivel de competencia. Además, estas pruebas afectan al conjunto de la sociedad y, por ello, deben participar muy directamente en este proceso de decisión los diversos colectivos implicados. Así, es conveniente llevar a cabo un proceso de determinación del estándar de tipo mixto, integrando los criterios absolutos con las consecuencias que de su aplicación se derivarían para diferentes colectivos.

De este modo, es muy importante tener en cuenta en este trabajo la minimización de los Errores de Selección, por lo que las labores de adecuación del Estándar como

Punto de Corte[13] son especialmente relevantes. La determinación de la puntuación de corte debe desarrollarse a partir de un proceso iterativo en el que se conjuguen técnicas de juicio con análisis empíricos y en el que el estudio de las consecuencias de aplicación del Estándar moderen las aplicaciones de juicio. En este contexto es básica la retroalimentación de información al Comité de Expertos que desarrolla las pruebas.

En estos procedimientos se tiende a seleccionar la puntuación de corte como aquella que maximiza la fiabilidad y minimiza los Errores de Selección (tipo I y tipo II). No obstante, atendiendo a la Razón de Pase[14], si es que está prefijada, el Comité de Expertos puede valorar la utilidad diferencial de asumir decisiones con Error tipo I o tipo II, de forma que se integren en la determinación de la puntuación de corte la composición de aquellos errores evaluativos que resulte menos lesiva para el adecuado uso del estándar.

Como en el caso de las pruebas anteriores estas pruebas deben sustentar su desarrollo como Modelo de Medida sobre la Teoría de Respuesta al Ítem. Sin embargo, en la selección de indicadores para el análisis de ítems y para la fiabilidad es preciso tener en cuenta la existencia del estandar criterial. Por ello, en la determinación de la fiabilidad deben contemplarse indicadores de Consistencia de la Decisión, de forma que pueda valorarse la capacidad de la prueba para diferenciar, al menos, entre sujetos que tienen y no tiene el nivel mínimo de competencia en el Dominio de Referencia. Por su parte, en el analisis de ítems se atenderá especialmente a la identificación de indicadores de discriminación, en los que deberán incluirse formulaciones que tengan en cuenta —además de la capacidad global de discriminación— la actuación consecuente con el Punto de Corte fijado.

• ***Pruebas de Nivel y de Dominio.***

Las pruebas de Nivel y las de Dominio las podemos considerar como variaciones de las anteriores, diferenciándose en virtud de la amplitud del Dominio Educativo a que se refieren —que es más reducido— o bien en relación a la amplitud de la población a la que van dirigidas —que tambien suele ser más específica—. Tienen en parte por tanto unos objetivos y características semejantes a las anteriores. En este caso, nos estamos refiriendo a pruebas que aporten información, por ejemplo, sobre si un estudiante ha alcanzado el nivel suficiente como para pasar de un curso a otro en una materia, o bien si ha superado los niveles mínimos exigidos en un programa de formación concreto[15] —por ejemplo, un programa de reciclaje para docentes sobre técnicas de observación en el aula—. El hecho de referirse a un programa educativo mucho más concreto conlleva diferencias sustanciales que se pueden sintetizar en las siguientes:

---

13 Entendemos por Punto de Corte aquella puntuación en la prueba que expresa el nivel mínimo de competencia. Proviene del Estándar y constituye el ajuste empírico del mismo, teniendo en cuenta criterios de fiabilidad (Jornet y Suárez, 1989b).

14 % de sujetos que pueden ser admitidos, por ejemplo, en un programa.

15 Se ha informado de algunos desarrollos específicos de pruebas —insertas en Modelos evaluativos— que podrían ser identificables en esta categoría como los trabajos de Rodríguez Lajo (1986), Jornet (1987) y Jornet et al. (1993).

a) se puede dar una definición del universo de referencia tanto empírica como teórica — esto último especialmente en los programa más concretos—,
b) se puede aspirar a la valoración del proceso y no sólo del producto. Al ser pruebas referidas a dominios más concretos cabe identificar unidades en el Dominio, a partir de las cuales se puedan realizar inferencias acerca de los procesos,
c) las reglas de conexión entre el universo y la prueba están mucho más determinadas —en numerosas ocasiones completamente determinadas—,
d) en el análisis de ítems, junto a los indicadores de los parámetros básicos resulta de interés (al ser pruebas de nivel o de evaluación de un programa específico) la sensitividad instruccional, como expresión de la capacidad de los ítems para discriminar las adquisiciones propias del programa,
e) la amplitud de las poblaciones de referencia puede condicionar el Modelo de Medida adecuado a cada caso. Así, se debe distinguir entre las pruebas que se orientan a poblaciones amplias y las que se destinan a ámbitos más concretos. Por ejemplo, hay que diferenciar entre una prueba de Cálculo para primero de Primaria aplicable a estudiantes valencianos y otra destinada a evaluar la competencia alcanzada en un programa de formación para la participación de Equipos Directivos de Centros. El Dominio se concreta en ambos caso, pero la primera situación va dirigida a una población amplia y le son aplicables los mismos referentes de Medida que los ya comentados en los casos anteriores, mientras que en la segunda situación las pruebas se deberán sustentar en la Teoría Clásica del Test y en indicadores que provienen del ámbito de la Evaluación Referida al Criterio[16].
f) el estándar de referencia está normalmente basado en un criterio absoluto —aunque se den, obviamente, casos en que se utiliza una combinación con información normativa—.

En cualquier caso, este tipo de pruebas se sitúa entre las pruebas amplio espectro y las de Aula, adoptando características que les son propias a los dos enfoques.

PRUEBAS DE CLASE O DE USO EN EL AULA

Las pruebas de Clase o pruebas de Aula hacen referencia a las que puede utilizar el profesor para la evaluación de sus alumnos. No obstante, es en este ámbito donde probablemente se aprecia una peor aceptación de las pruebas estandarizadas. Y ello, porque se atribuye a la Estandarización condiciones que alejan estas pruebas de la individualización.

Sin embargo, debe tenerse en cuenta que ambos —Estandarización e Individualización— no son conceptos necesariamente contradictorios; más bien, es el tipo de uso que se realice de las pruebas lo que puede enfrentarlos. La Estandarización mejora

16 En ambos casos se pueden utilizar aportaciones derivadas de la Teoría de la Generalizabilidad para conseguir indicadores más consistentes de los parámetros de la prueba, especialmente en aquellos casos en los que no sea adecuado utilizar los modelos TRI (Brennan, 1983; Shavelson y Webb, 1991).

esencialmente las condiciones de objetivación de la medida. Y ello a veces se olvida por parte de los detractores de las pruebas, poniendo el énfasis sobre aspectos propios de la individualización que, sin embargo, podrían se atendidos con pruebas estandarizadas, sin considerar que los procedimientos no-estandarizados no resuelven adecuadamente los problemas derivados de la subjetividad del observador o el evaluador.

Por otra parte, el marco derivado de la LOGSE ha puesto de manifiesto la necesidad de individualizar o personalizar los diseños curriculares. Un problema práctico al que habitualmente aluden los profesores es la falta de metodología adecuada para el desarrollo de las Adaptaciones Curriculares Individualizadas. Obviamente, las soluciones no están sólo en la Medida y la Evaluación, pero éstas constituyen un elemento instrumental inicial desde el que abordar la solución de este problema.

Así, debe tenerse en cuenta que el desarrollo de pruebas estandarizadas para la evaluación de una unidad didáctica es probablemente el marco donde puede disponerse de pruebas más válidas. Esto es así, dado que los Dominios Educativos de referencia en estos casos constituyen Universos de Medida finitos, claramente especificables, concretos.

Este hecho afecta directamente a la Validez de Contenido, pudiendo aspirar en este contexto a pruebas más representativas del Dominio Educativo del que se derivan. Por otra parte, si se estructura el Dominio Educativo de forma perfectamente asociada al planteamiento metodológico-didáctico, la Validez de Constructo también puede verse beneficiada. Esto es así, no sólo por el hecho de la asociación trabajo de aula-sistema de evaluación (que sería una expresión más concreta de la Validez de Constructo, como Validez Curricular), sino muy especialmente por las características del desarrollo del aprendizaje, en el que se podrá reflejar el constructo teórico que esté a la base del diseño curricular y del enfoque metodológico-didáctico del programa.

Además, los Dominios Educativos, en estos casos, se refieren a unidades didácticas —o lecciones— por lo que incluyen pocas unidades, lo que favorece el micro-análisis de todas las tareas-ítems implicados en el Dominio. Pueden permitir, pues, una definición exhaustiva de la población de conductas que pongan de manifiesto las adquisiciones (habilidades, capacidades y destrezas) a que hace referencia un Dominio. De esta forma, en la definición de este tipo de Dominios cabe identificar unidades de medida, a partir de las cuales se puedan inferir interpretaciones procesuales bastante precisas, así como de productos específicos.

Ello favorece que este tipo de pruebas pueda estar muy bien adaptado para la medición y evaluación de procesos y productos de aprendizaje[17]. De este modo, las posibilidades de interpretación se abren: no sólo se puede interpretar un nivel de competencia —a partir de la puntuación total— sino explicar el nivel, informando de los procesos de adquisición —a partir de la interpretación particular de los ítems—.

Este hecho se ve favorecido porque la situación de Medida que puede darse en un Aula no tiene por qué limitarse a una situación de examen habitual tipo test. En este

17 Recientemente se ha venido informando de propuestas de interés en nuestro ámbito educativo, como las de Buendía y Salmerón (1994) o las de Toboso (1995 a y b).

marco, pueden formar parte de la "prueba" diversos tipos de elementos: desde ítems clásicos de lápiz y papel hasta ítems micro-situacionales en los que la valoración provenga de la observación de la tarea que realiza cada individuo.

La definición del Dominio, establecimiento de Reglas de generación de ítems y escritura de ítems, la realiza —como en casos anteriores— un comité de expertos, pero en esta situación, está compuesta por los profesores de una materia (o departamento, o equipo de ciclo).

En el trabajo a realizar por el Comité hay que tener en cuenta los siguientes aspectos:

- Antes de desarrollar la prueba es esencial definir el rol que ésta tendrá dentro de los recursos evaluativos de que disponga el profesor. El contexto ideal de uso es aquel en el que se identifican fuentes múltiples y diversos instrumentos.
- En el desarrollo del Análisis del Dominio la reflexión deberá orientarse hacia elementos de relevancia de los ítems respecto de los objetivos que pretenden medir (Congruencia Ítem-Objetivo) así como respecto a la representatividad de los ítems —como situación evaluativa— en relación a los planteamientos metodológico-didácticos seguidos en el Aula.
- Aunque en este contexto no resulta tan trascendente el análisis del Sesgo de los ítems, su anticipación —por procesos de juicio— es una labor de especial interés. Así, junto a variables básicas como Sexo o Lengua —en Comunidades bilingües—, dependiendo del nivel educativo, pueden producirse sesgos en situaciones de apertura del currículum. De esta manera, pueden haber alumnos reforzados en su aprendizaje de una materia por el efecto del aprendizaje de otras opcionales. Este tipo de sesgo sería importante identificarlo a efectos de determinar adecuadamente el nivel de competencia a que puede aspirarse en la materia en la que se desarrolla la prueba.
- No obstante, aunque existan sesgos de los que necesariamente tengamos que independizar a las pruebas, otros —como el último mencionado— puede ser difícil de eliminar, por lo que al menos es importante identificarlos, conocerlos, y asignarles valor diagnóstico o modulador de las decisiones evaluativas.
- El análisis del sesgo en este caso se basa más sobre procesos cualitativos de juicio que sobre la comprobación empírica de los resultados obtenidos por las personas evaluadas. El factor clave para poder llevar a cabo una comprobación adecuada es el tamaño de la muestra, en estos casos muy reducida.

Respecto a los componentes técnicos derivados del Modelo de Medida, en este contexto no puede utilizarse la Teoría de Respuesta al Ítem, dado que el tamaño de los grupos que se trabaja es muy reducido.

No obstante, dentro de la Teoría Clásica de los Tests se dispone de indicadores suficientes que pueden, realizando las adecuadas adaptaciones en su uso e interpretación, operacionalizar los análisis necesarios. Asimismo, en el marco de la Evaluación Referida al Criterio existen múltiples indicadores de fácil utilización que racionalizan la lógica de selección criterial.

Sin embargo, hay que tener presente que las características derivadas del objeto de

medida y del uso de la prueba, así como las que devienen de los tipos de distribución que podemos encontrar en pequeñas muestras[18], conllevan necesarias modificaciones en la utilización e interpretación de indicadores clásicos. De esta manera, características básicas empíricas útiles en la selección de ítems del Modelo Clásico no serían deseables aquí. El valor de los parámetros e indicadores radica no tanto como elemento de selección de los ítems (que se sustentará preferentemente sobre procesos de juicio) como elemento de información para el grupo de profesores —Comité— que desarrolla la prueba. Su valor como indicadores de selección de ítems, se circunscribe a la comprobación de las hipótesis funcionales que el Comité haya anticipado para los ítems, respecto a la dificultad teórica y su capacidad de discriminación primordialmente (Jornet y Suárez, 1994 ; Rivas, Jornet y Suárez, 1995).

¿Qué tipo de interpretaciones se requieren?

En este contexto, en donde se deben tomar decisiones acerca de la promoción de cada persona en su aprendizaje, es preciso conocer adecuadamente su posición respecto del Dominio Educativo, por lo que la interpretación necesariamente debe ser criterial, basada en un estándar absoluto.

Para el desarrollo del Estándar y su especificación como puntuación de corte la aportación del Comité de Expertos es nuevamente esencial. Entre los métodos en que pueden apoyarse, en este contexto de desarrollo de pruebas, cobran especial relevancia aquéllos que utilizan como información para retroalimentar al Comité en su proceso de determinación, el análisis de las consecuencias de su aplicación sobre sujetos conocidos (Livingston y Zieky, 1982). Estos usos, en la práctica, se convierten en evidencias de validación de la puntuación de corte.

En cualquier caso, los procesos de evaluación en el Aula hay que considerar que no se debe aspirar a que se sustenten sobre pruebas estandarizadas exclusivamente. La oportunidad de su utilización depende fundamentalmente de que sean adecuadas al tipo de materia que se pretenda evaluar. La defensa del uso de este tipo de pruebas para la evaluación debe realizarse desde el marco en que realmente sean más útiles (fiables y válidas) que otras alternativas o técnicas evaluativas. Asimismo, debe tenerse en cuenta que las pruebas estandarizadas deben derivarse desde un programa educativo bien establecido y, como instrumento están al servicio del mismo. Un peligro genérico que nace del uso de cualquier sistema de evaluación es que acaben siendo los instrumentos los que constituyan la referencia para el desarrollo del programa, acabando por condicionar su uso (De la Orden, 1993).

PRUEBAS DE PROPÓSITO DIAGNÓSTICO

Las ***Pruebas de Propósito Diagnóstico*** rompen con el discurso de lo más general a lo más concreto que relaciona los tres tipos que hemos expuesto. Así, mientras en las pruebas revisadas con anterioridad el objetivo es, en términos generales, valorar las

18 No sería aceptable anticipar como efecto educativo que el Rendimiento se distribuirá como la Curva Normal. Es más lo habitual —y deseable— sería que los efectos educativos fueran asimilables a distribuciones beta, con tendencia asimétrica negativa.

adquisiciones sobre un programa educativo, más o menos amplio, aquí se trata de poder indagar respecto a la posible existencia de determinados problemas de aprendizaje y cuáles son sus características concretas. El objetivo es, pues, delimitar si se da un determinado problema en el proceso normal de adquisiciones que desarrolla un sujeto y poder extraer información sobre la cualidad de tal problema, de modo que se pueda orientar mejor la subsiguiente intervención.

En este grupo incluimos dos grandes tipos de pruebas: referidas al Currículum y de diagnóstico propiamente dichas.

Las pruebas estandarizadas referidas a un currículum tienen por objeto indagar acerca de la posición de un sujeto respecto a un Diseño Curricular dado (que actúa como Dominio Educativo). Las hemos clasificado aquí porque su propósito generalmente es de tipo diagnóstico, dado que se trata de recabar información independiente de las calificaciones escolares —y sin finalidad de uso en el contexto del Aula— acerca de si los sujetos tienen adquisiciones básicas correspondientes con su desarrollo curricular-escolar, o bien presentan disfunciones—[19].

Generalmente, estas pruebas pretenden abarcar Dominios amplios, correspondientes a dimensiones que se identificables a través de un Diseño Curricular de largo alcance (como por ejemplo, Numeración, Cálculo Mental, Interpretación de Datos o Resolución de Problemas). Sin embargo, en el análisis del Dominio prevalece la identificación de los elementos clave que se asocian a diferentes etapas de adquisición. Esta identificación es la que permite situar al sujeto en su nivel de aprendizaje en el Dominio.

Por su parte, las pruebas Diagnósticas propiamente dichas tienen como finalidad no sólo determinar la posición del sujeto en el Dominio de referencia, sino describir adecuadamente los elementos deficitarios con el fin de planificar la intervención. En este sentido, las unidades del Dominio deben estar claramente definidas y previamente analizadas respecto a su asociación con unidades de intervención.

Es habitual que en estas situaciones se parta de un marco teórico que define dimensiones respecto de las cuales su puede concentrar la información significativa para la toma de decisiones, aunque no siempre tenga la consistencia teórico-metodológica debida (De la Orden et al. 1994).

Como señala Oosterhof (1994), las pruebas de propósito diagnóstico deben ser utilizadas con cautela, pues la investigación básica acerca de los constructos en ellas implicados, todavía es escasa[20].

En relación directa con esto, el objetivo de la evaluación con estas pruebas es mixto, en el sentido que pueden estar orientadas al proceso educativo, al producto o a cualquier combinación de ambos objetivos. Esto conlleva que las pruebas estén basa-

---

19 Aunque no son muy frecuentes, existen ya algunas pruebas de interés como por ejemplo la Escala Key-Math R de Connolly (1988), adaptada por Marí (1996) a nuestro contexto educativo o algunas otras desarrolladas directamente en el mismo, como la Batería de Pruebas de Lenguje FCI (Bartolomé, et al., 1985).

20 Sin embargo, se pueden identificar ya desarrollos muy adecuados (como la prueba CRIL de Lenguaje de Wiig, 1990, de la que parte el desarrollo de la prueba ICL de Puyuelo y Renom —1993— y Puyuelo, Renom y Solanas —1995—).

das en unas reglas de conexión con el universo tan específicas como lo permita las características y la amplitud del mismo. Finalmente, el estándar en que se basa la decisión suele ser una combinación de indicadores absolutos y normativos. Esto es así dado que para la determinación de la existencia de un problema suele ser tan útil emplear definiciones absolutas que reflejen las claves de su identificación (nivel en que se produce un problema, patrón procesual del mismo, etc...) como la información relativa al grupo de pertenencia para situar la dimensión característica del mismo —por ejemplo, un problema de inversiones en la lectura dependerá tanto de una determinada frecuencia concreta como de la situación relativa dependiendo del grupo de edad al que pertenece el sujeto—.

PRUEBAS INDIVIDUALIZADAS

En este caso el objeto es proporcionar un sistema de recogida de información muy flexible que se ajuste a las características de cada sujeto o situación de medida y que proporcione, por ello, una información más rica y significativa en los puntos críticos. Como se aprecia en el Cuadro 2, no siempre la unidad de referencia es un sujeto concreto, pudiendo ser un currículum o programa completo. Además, la adaptación puede realizarse de forma estática o dinámica. En el primer caso la prueba entera se construye en función de las características o directrices del grupo o situación de referencia, mientras que en el segundo caso es el propio rendimiento el que proporciona el patrón de referencia para la adaptación sucesiva de la prueba.

Estamos hablando de pruebas que precisan de una definición lo más exhaustiva posible del Dominio Instruccional de referencia y de unas reglas sumamente concretas de relación entre el Universo Instruccional y la prueba. De no producirse estas condiciones no se podría establecer bien el ajuste para cada situación o individuo.

En general, las pruebas adaptadas se fundamentan en bancos de objetivos e ítems que ofrecen algunas grandes organizaciones públicas y privadas. Los dos formatos esenciales se dan en función de que sea la propia empresa u organización que facilite la adaptación ya completada al usuario final o que se le suministre la información y una herramienta informática a este último para que pueda hacer esta adaptación por sus propios medios. Así, los sistemas AIMS (Academic Instructional Measurement Systems) de The Psychological Corporation, ORBIT (Objective-Referenced Bak of Ítems and Tests) de CTB/McGraw-Hill o MULTISCORE de The Riverside Publishing Company están compuestos por unos centenares a miles de objetivos y muchos miles de ítems que abarcan la mayor parte de los ámbitos curriculares de la educación primaria y secundaria.

La adaptabilidad en el caso de situaciones o programas dependerá de la cantidad de opciones disponibles sobre el dominio (objetivos, ítems, etc.), de las informaciones sobre la estructura y características métricas de los elementos (dificultad, discriminación, elección de alternativas, recomendaciones asociadas, etc.) y de la existencia de mecanismos para integrar variaciones propias de cada situación en la prueba (herramientas que permitan el desarrollo de objetivos, ítems con diversas variantes, recomendaciones, etc.).

En el caso de los tests con adaptación instantánea a la ejecución por el sujeto —como el MicroCAT Testing System (Assessment Systems Corporation, 1988) o los WICAT Systems (1986)— sirven esencialmente las ideas que acabamos de apuntar. No obstante, la adaptabilidad en estos casos se incrementa cuando esta pruebas se puede adaptar mejor al sujeto en la situación específica de aplicación que sirve como referencia (sea el programa, el sistema específico de recuperación, etc.). Asimismo, la información que se puede obtener en estas situaciones es tan rica que difícilmente se puede encarar una prueba de este tipo solamente en función de una valoración del producto. No obstante, este es el enfoque predominante todavía, pues se ha producido una adaptación excesivamente mimética respecto a las pruebas tradicionales. Piénsese que estamos hablando de pruebas que por su complejidad, normalmente, precisan de un soporte de tecnología informática, ya que el único medio que facilita una gran precisión y variedad en la recolección de información. De hecho, este tipo de medida es el horizonte natural de los sistemas EAO al incorporar la evaluación dinámica que se precisa en estos casos. Hoy en día, con todo, la mayoría de los sistemas EAO están lejos de adoptar las posibilidades de medida inherentes a este tipo de pruebas —de hecho, tienen serios problemas para cubrir las mínimas directrices que garanticen una valoración del rendimiento—. Existen, no obstante, algunas excepciones que constituyen caminos muy sugerentes, como la experiencia del Cognition Technology Group en la Universidad de Vanderbilt (Goldman, Pellegrino y Bransford, 1994), con planteamientos integrales de enseñanza y evaluación-medida que pueden aportar respuestas a algunas necesidades de transformación que ya hemos señalado.

La calidad de la Definición del Dominio de estas pruebas basadas en bancos de ítems es bastante elevada[21]. En cualquier caso, el nivel de especificación que requieren permite una valoración ajustada y actualizada de estos sistemas que sirva de referencia para nuestra actuación. Por ejemplo, se realizan revisiones de estos sistemas periódicamente que nos añaden referentes de validez y utilidad de los mismos (Naccarato, 1988).

Los Modelos de Medida asociados a las aplicaciones más consistentes están basados en Teoría de Respuesta al Ítem, tanto para la composición del banco de referencia como para su gestión en cada situación de evaluación concreta (Kingsbury y Zara, 1989). No obstante, como ya hemos señalado, el campo es muy heterogéneo y desestructurado, hallándose incluso pruebas que no están soportadas por modelo alguno de medida. En definitiva, para los proyectos de una cierta envergadura —respaldados por instituciones o empresas de suficiente solvencia— las herramientas disponibles en los modelos TRI constituyen la recomendación universalmente aceptada. Las limitaciones son las mismas que ya se han descrito respecto a otras pruebas y las ventajas son aún mayores, al entramarse los procedimientos con mayor facilidad en una estructura de aplicación basada en ordenador. Queda por resolver, a pesar de todo, una amplia variedad de temáticas y situaciones para las que, al igual que en otros contextos de evaluación, no cesan de proponerse alternativas parciales que, siendo muy

21 Otra cuestión bien diferente la constituyen los sistemas de evaluación ligados a las aplicaciones EAO, cuya calidad metodológica general es muy desigual.

CUADRO 2
SÍNTESIS DE PROCEDIMIENTOS DE PRUEBAS INDIVIDUALIZADAS

Tipos de Prueba o procedimiento	Objetivo	Características de las Tareas/Pruebas
PRUEBAS DE ADMINISTRACIÓN INDIVIDUAL	Mejorar la precisión en la estimación de la habilidad del sujeto	Las tareas están graduadas, en ocasiones se basa la administración en senderos de decisión.
PRUEBAS DE FORMAS MÚLTIPLES	Obtener múltiples medidas paralelas, equivalentes —o al menos comparables— de cada sujeto.	Pruebas estadísticamente paralelas
PRUEBAS ESTANDARIZADAS DE NIVELES MÚLTIPLES	Simplificar la medida ajustándola al nivel de habilidad del sujeto	Items basales; determinaciones del nivel inicial de partida de la prueba.
TESTS ADAPTATIVOS • DE NIVEL	Maximizar la precisión de la estimación de la habilidad de cada sujeto, utilizando el mínimo número de ítems	Ítems clave-característicos de niveles.
DIAGNÓSTICOS	Diagnosticar las dificultades de aprendizaje concretas del sujeto	Selección específica de los ítems a administrar en función del nivel inicial demostrado por el sujeto en ítems de prueba.
SOPORTES TECNOLÓGICOS A LA INVIDUALIZACIÓN • BANCOS DE ÍTEMS TRADICIONALES	Automatizar la construcción de pruebas a partir de una definición genérica del Dominio, con ítems cerrados e identificados por sus parámetros.	Pruebas ajustadas a diseños curriculares y adaptadas a niveles específicos.
• BANCOS DE REGLAS DE GENERACIÓN DE ÍTEMS	Automatizar la construcción de ítems, y en algunos casos, incluso su administración. Generar múltiples pruebas paralelas.	Tests paralelos en contenido, diferentes para cada sujeto y que pueden ajustarse a los diversos curricula.

sugerentes, no se han estructurado en ningún planteamiento suficientemente sólido hasta el momento presente.

Por lo que respecta a los estándares, aunque también utilizan información respecto a criterios diferenciales, lo fundamental de las decisiones consiste en las definiciones absolutas que incorporan. De hecho, en muchos casos los referentes diferenciales son parciales o imposibles, en función de la adaptación que se realice —por ejemplo, si se añade una cantidad sustancial de modificaciones en la definición de algún subdominio educativo todo referente normativo a este respecto carecería de validez—. En otras

situaciones, la información normativa solamente puede actuar como referente marco relativamente alejado, dado que el propósito de estas pruebas suele ser más formativo u orientado a la recuperación. Desafortunadamente existe poco trabajo hecho en este ámbito e incluso las orientaciones reflejadas a este respecto en las "normas y orientaciones de actuación" (APA, 1986) no parecen haber madurado suficientemente la necesaria adaptación a estas situaciones de medida-evaluación.

ALGUNAS NOTAS FINALES

La evolución de los métodos de construcción de pruebas aporta una base bastante sólida para el desarrollo de instrumentos estandarizados de evaluación. Para nuestro ámbito educativo, los desarrollos son muy escasos, aunque crecientes, en consonancia con la progresiva implantación de actividades de evaluación. El arraigo de una cultura evaluativa sin duda conllevará la necesidad de utilizar instrumentos mucho mejor construidos y adaptados que los que habitualmente se utilizan. La demanda de calidad también afectará a los instrumentos evaluativos. Sólo será posible responder a estos retos si abordamos decididamente el desarrollo de instrumentos de medida educativa, que respondan a las necesidades de los diversos programas y fenómenos a evaluar.

La institucionalización de la revisión del sistema educativo y de los diferentes componentes y actores del mismo es una realidad a la que necesariamente se debe responder con instrumentos mejor diseñados. Incluso, hechos evaluativos que afectan muy directamente a nuestra sociedad —como es la selectividad universitaria— en la actualidad aún se desarrolla sobre esquemas imprecisos, que hacen que ésta no responda en definitiva al sistema y que no se pueda hablar de equidad evaluativa. Actuaciones como la selectividad están reclamando respuestas profesionales evaluativas que, al menos, integren las opciones metodológicas disponibles.

En el campo del Diagnóstico Educativo también es evidente la carencia de instrumentos. De hecho no se dispone de Baterías a lo largo de los diferentes niveles y áreas educativas que cubran las dimensiones esenciales del Rendimiento. Otro tanto puede decirse respecto a las versiones individualizadas, como las Pruebas Asistidas por Ordenador, que además deben reivindicar su existencia frente a los exiguos sistemas de evaluación que incorporan las aplicaciones de Enseñanza Asistida por Ordenador.

Respecto a las evaluaciones en el Aula deben estar soportadas sobre una profunda reflexión por parte de los profesores acerca de los componentes de sus diseños curriculares. Un aspecto central de esta reflexión debe ser el sistema de evaluación. El desarrollo de instrumentos no tiene por qué ser la finalidad, pero sí constituye un buen medio de revisión de los componentes de un programa educativo, así como de los elementos que inciden en su realización. Incorporar elementos propios de las pruebas estandarizadas, como son el análisis de los Dominios Educativos o el Desarrollo de Estándares, aunque no se persiga ni se llegue a una estandarización completa, supone indudablemente integrar elementos de mejora de los procesos evaluativos. Obviamente, en muchas ocasiones se afirma que el profesorado no ha sido formado para abordar este tipo de procesos. La respuesta es clara: es necesario reforzar la

formación de estos profesionales en estas áreas, dado que son instrumentales para su actuación docente.

Por último, la estandarización de la medida, como base para la evaluación, si quiera en sus versiones más "tecnológicas" —como las Pruebas Asistidas por Ordenador—, no implica necesariamente un empobrecimiento de la información necesaria en la evaluación. Existen suficientes vías de trabajo para incorporar definitivamente la medida de tareas complejas, el proceso de construcción de los conocimientos, el aprendizaje cooperativo o el pensamiento crítico. En buena medida, nos tenemos que comprometer en realizar un esfuerzo por avanzar en esas direcciones y tratar de situarnos en línea con los países más desarrollados en estos ámbitos disciplinares.

REFERENCIAS BIBLIOGRÁFICAS

AMERICAN PSYCHOLOGICAL ASSOCIATION (1986) *Guidelines for computer-based tests and interpretation.* Washington, D.C.: Autor.

ANGOFF, W.H. (1988) Validity: An evolving concept. En H. WAINER y H.I. BRAUN (Eds.) *Test Validity.* Hillsdale, NJ: LEA.

BARTOLOMÉ, M.; BISQUERRA, R.; CABRERA, F.; ESPÍN, J.V.; MATEO, J. Y RODRÍGUEZ, Ll. (1985) *Batería de Pruebas de Lenguaje Final de Ciclo Inicial.* Barcelona: CEAC.

BRENNAN, R.L. (1983) *Elements of Generalizability Theory.* Iowa City, IA: American College Testing Program.

BUENDÍA, L. y SALMERÓN, H. (1994) Construcción de pruebas criteriales de aula. *Revista de Investigación Educativa, 23,* 405-410.

CERI/INES (1995) *Education at a Glance.* OECD Indicators. París: OECD.

CONNOLLY, A.J. (1988) *Key Math Revised: a diagnostic inventory of essential mathematics.* Circle Pins, Minnesota: American Guidance Service.

DE LA ORDEN, A. (1993) Influencia de la evaluación del aprendizaje en la eficacia de la enseñanza. *Revista de Investigación Educativa, 22,* 7-42.

DE LA ORDEN, A.; GAVIRIA, J.L.; FUENTES, A. y LÁZARO, A. (1994) Modelos de construcción y validación de instrumentos diagnósticos. *Revista de Investigación Educativa, 23,* 129-178.

DE MIGUEL, M. et al. (1994) *Evaluación para la calidad de los Institutos de Educación Secundaria.* Madrid: Escuela Española.

GIL, G.; GONZÁLEZ, A. y SUÁREZ, J.C. (1995) Un modelo de construcción de pruebas de rendimiento para la evaluación de las enseñanzas mínimas en la Educación Primaria. En AIDIPE (Comp.) *Estudios de Investigación Educativa en Intervención Psicopedagógica.* Valencia: AIDIPE.

GOLDMAN, S.R., PELLEGRINO, J.W. y BRANSFORD, J.D. (1994) Assessing programs that invite thinking. En E. BAKER y H.F. O'NEIL Jr. (Eds.) (1994) *Technology Assessment in Education and Training.* Hillsdale, NJ: LEA.

HALADYNA, T.M. (1992) Context dependent ítem sets. *Educational Measurement: Issues and Practice. 11,* 11-25.

HAMBLETON, R.K. (1984) Validating the tests scores. En R. BERK (De.) *A guide to*

Criterion-Referenced Tests construction. Baltimore, Mass.: Johns Hopkins University Press.

HAMBLETON, R.K.; SWAMINATHAN, H. (1985) *Ítem Response Thory: Principles and Applications*. Norwell, MA: Kluwer.

INCE (1996) *Lo que aprenden los alumnos de 12 años. Evaluación de la Educación Primaria. Datos básicos. 1995*. Madrid: Centro de Publicaciones del Ministerios de Educación y Cultura.

JORNET, J.M. (1987) *Una aproximación teórico-empírica a los métodos de medición de referencia criterial*. Tesis Doctoral. Valencia: Universitat de València.

JORNET, J.M. y SUÁREZ, J.M. (1989a): «Conceptualización del Domínio educativo desde una perspectiva integradora en Evaluación Referida al Criterio». *Bordón. 41*, 2, 237-275.

JORNET, J.M. y SUÁREZ, J.M. (1989b): «Revisión de Modelos y Métodos en la determinación de estándares y en el establecimiento de un Punto de corte en Evaluación Referida al Criterio (ERC)». *Bordón. 41*, 2, 277-301.

JORNET, J.M. y SUÁREZ, J.M. (1994) Evaluación Referida al Criterio. Construcción de un Test Criterial de Clase. En V. GARCÍA HOZ (Dir.) *Problemas y Métodos de Investigación en Educación Personalizada*. Madrid: Rialp.

JORNET, J.M., SUÁREZ, J.M., GONZÁLEZ SUCH, J., PÉREZ CARBONELL, A. y FERRÁNDEZ, M.R. (1993) *Evaluation Report of the Project: Communication and Presentation Skills for Technological Transfer Agents*. Euro-Innovations-Manager. Valencia: ADEIT/IMPIVA/CEEI.

KINGSBURY, G.G. y ZARA, A.R. (1989) Procedures for selecting ítems for computerized adaptive testing. *Applied Measurement in Education 2(4)*, 359-375.

LADUCA, A., STAPLES, W.I., TEMPLETON, B. y HOLZMAN, G.B. (1986) Ítem modelling procedure for constructing content-equivalent multiple-choice questions. *Medical Education, 20*, 53-56.

LINN, R.L. (1990) Has Ítem Response Theory increased the Validity of Achievement Test scores? *Applied Measurement in Education, 3*, 2, pp. 115-141.

LIVINGSTON, S.A. y ZIEKY, M.J. (1982) *Passing Scores*. Princeton N.J.: ETS.

MARI, R. (1996) *Evaluación del Rendimiento en Matemáticas: adaptación de la Escala Key Math-R*. Tesis Doctoral (en prensa: microficha). Valencia: Universitat de València.

MARTÍNEZ RIZO, F. (1996) *La calidad de la educación en Aguascalientes. Diseño de un sistema de monitoreo*. Aguascalientes, México: Universidad Autónoma de Aguascalientes (UAA)-Instituto de Educación de Aguascalientes (IEA).

NACCARATO, R.W. (1988) *A guide to item banking in Education* (3ª ed.) Portland, O.: Northwest Regional Education Laboratory.

ORDEÑANA, B. (1991) Funcionamiento diferencial de los ítems: una aplicación al campo de las diferencias entre sexos. *Revista de Investigación Educativa, 9*, 17, 119-128.

OSTERLIND, S.J. (1992) *Constructing test items*. (2ª ed.). Boston: Kluwer.

PÉREZ JUSTE, R. y MARTÍNEZ ARAGÓN, L. (1989) *Evaluación de centros y calidad educativa*. Madrid: Cincel.

POSTLETHWAITE, T.N. (1987) Introduction: Special issue on the Second IEA Study. *Comparative Educational Review. 31*(1), 150-158.

RECKASE, M.D. (1979) Unifactor latent trait models applied to multifactor tests. *Journal of Educational Statistics. 4,* 207-230.

RIVAS F. et al. (1986): *Proyecto Valencia: Objetivos básicos de aprendizaje en los Ciclos y Areas de Lenguaje y Matemáticas en la EGB. Una aproximación de Evaluación Referida al Criterio.* Valencia: Servicio de Estudios y Publicaciones Universitarias, S.A.

RIVAS, F. JORNET, J.M. y SUÁREZ J.M. (1995) Evaluación del aprendizaje escolar: claves conceptuales y metodológicas básicas. En F. SILVA (De.): *Evaluación psicológica en niños y adolescentes.* Madrid: Síntesis.

RODRÍGUEZ LAJO, M. (1986) Evaluación del rendimiento criterial vs. Normativo. Modelo de evaluación FCO. *Revista de Investigación Educativa. 3,* 6, 304-321.

ROID, G.H. y HALADYNA, T.M. (1982) *A technology of test-item writing.* New York : Academic Press.

SAMEJIMA, F. (1974) Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika. 39,* 111-121.

SHAVELSON, R.J. y WEBB, N.M. (1991) *Generalizability Theory. A Primer.* Newbury Park, CA.: SAGE.

SPECIAL STUDY PANEL ON EDUCATION INDICATORS (SSPEI) (1991) *Education counts. An indicator system to monitor the nation's educational health.* Washington : National Center for Educational Statistics. USA Department of Education.

TEJEDOR, F.J.; GARCÍA VALCÁRCEL, A. y RODRÍGUEZ CONDE, M.J. (1994) Perspectivas metodológicas en la evaluación de programas en el ámbito educativo. *Revista de Investigación Educativa. 23,* 93-128.

TOBOSO J. (1995a): Fundamentos teóricos del proceso evaluador desde el marco curricular de la LOGSE. En AIDIPE (Comp.): *Estudios de Investigación Educativa en Intervención Psicopedagógica.* Valencia: AIDIPE.

TOBOSO J. (1995b): Estudio empírico sobre la Evaluación de componentes cognitivos en la Resolución de problemas. En AIDIPE (Comp.): *Estudios de Investigación Educativa en Intervención Psicopedagógica.* Valencia: AIDIPE.

TOCA, M.T. y TOURON, J. (1989) Factores del Rendimiento Académico en los Estudios de Arquitectura. *Revista de Investigación Educativa, 7,* 14, 31-47.

TOURON, J. (1985) La predicción del rendimiento académico: procedimientos, resultados e implicaciones. *Revista Española de Pedagogía, 169-170,* 473-495.

WEISS, D.J. y YOES, M.E. (1991) Item Response Theory. En R.K. HAMBLETON y J.N. ZAAL (eds.) *Advances in Educational and Psychological Testing.* Boston, MA: Kluwer

WEST, A.; PENNELL, H.; THOMAS, S. y SAMMONS, P. (1995) Educational performance indicators. *EERA Bulletin, 1,* 3, 3-11.